Human Computer Interaction - Evaluation

Michael Granitzer Christin Seifert

Chapter HCI:IV

Chapter HCI:IV

I. Introduction

- Motivation
- Evaluation Planning
- Overview of Methods

Introduction

Motivation

Extensive UI evaluation is not an option, it is a necessity!

Benefits of UI evaluation:

- Identify the behaviour of users (they might not behave as expected)
- Quantify the usability of a certain UI design
- Find unexpected interaction problems (unsolvable tasks)
- Get directions for UI refinement
- Compare different UI designs
- Fullfill contracts (and have appropriate documentation)
- Ensure that users like the software for which developers spend time and effort

Introduction

Evaluation Planning

Multiple factors define when a UI should be evaluated:

- Stage of design (early, middle, late)
- Novelty of the project (well defined vs. exploratory)
- Number of expected users
- Criticality of the user interface (surgery support system vs. database administration tool)
- Costs of the product and finances allocated for testing
- Available time
- Experience of design and evaluation team

Practical dimensions:

- **Users:** from 3 users to hundreds of users
- **Time:** from one-time testing to muliple test phases over 2-3 years
- **Costs:** from 5% to 20% of the overall project budget
- Focus on task coverage rather then number of users

Introduction

Evaluation Methods

- **Expert Reviews:** the evaluator is part of the product team (colleagues, consultants), not the actual user
- **Usability Testing:** formal testing and observation of users in controlled environment
- **Surveys:** questionaires for users
- **Acceptance Tests:** testing againts requirements, outside lab environment
- **Evaluation During Active Use:** interviews, data logging, after product shipment

Chapter HCI:IV

II. Expert Reviews

- Overview
- Heuristic Evaluation
- Guidelines Review
- Consistency Inspection
- Cognitive Walkthrough

Expert Reviews

Overview

Expert Reviews comprise a set of usability evaluation techniques that are performed by usability and/or domain experts.

- UI design evaluated by experts, domain experts and UI design experts
- efficient way to identify usability problems
- a small group of usability experts finds 75% of all usability problems
- should not be used exclusively, because experts are not the real users
- typically performed in-house, fast and cost-efficient

Expert Reviews

Heuristic Evaluation

Heuristic Evaluation is performed by expert reviewing the UI according to a list of usability heuristics. The chosen heuristics depend on the interface type (e.g. different heuristics for medical applications and computer games).

Procedure: A checklist with the usability heuristics has to be prepared in advance. The evaluator then uses the user interface and fills out the checklist, optionally screenshots are taken when problems are encountered. The evaluation takes place in two passes: first pass focuses on general flow, second on particular elements in detail. Rate the found problems according to their severity. Only a fraction of all problems is found by a single evaluator, so use several evaluators (in practice 3-4) and aggregate the findings.

Outcome:

- sorted list of usability issues (list of violations of the heuristics used)

Expert Reviews

Guidelines Review

Guidelines Review (or Guideline Checking) is a usability inspection method and reviewing technique. An evaluator checks the UI againts a list of guidelines. Guideline documents can contain thousands of guidelines, thus the understanding of the document and the actual review can take considerable time. Guideline reviews can be performed early in the development process.

Outcomes:

- a list of violated guidelines

Non-Outcomes:

- quantitative analysis of the user interface

Extension Guideline Scoring: The list of guidelines is weighted according to their expected importance. The outcome of the walkthrough then is a score of non-conformaty of the UI.

Expert Reviews

Cognitive Walkthrough

Cognitive Walkthrough is a usability inspection method and reviewing technique. It is not done by real users but by evaluators playing the roles of users. It also does not require an executable of the interface, i.e., it can be performed using a real user interface, user interface mock-up or even a paper prototype.

Procedure: The user tasks described in the use cases or in the requirements definition are analysed in detail. For each task a goal and subgoals are identified. Further it is defined when the goals and subgoals are considered to be accomplished. This is informally called the "happy path" - a sequence of actions performed by the users in order to accomplish the task. Then the evaluator performs the walkthrough using the user interface answering predefined questions for each identified subgoal.

Expert Reviews

Cognitive Walkthrough

Questions to answer during the walkthrough:

1. Will the user try to achieve the effect that the subtask has? Does the user understand that this subtask is needed to reach the user's goal?
2. Will the user notice that the correct action is available? E.g. is the button visible?
3. Will the user understand that the wanted subtask can be achieved by the action? E.g. the right button is visible but the user does not understand the text and will therefore not click on it.
4. Does the user get feedback? Will the user know that they have done the right thing after performing the action?

Outcomes:

- fine-grained scenario description (outcome of the preparation phase)
- list of missing features
- action list for development

Non-Outcomes:

- quantitative analysis of the user interface

Expert Reviews

Consistency Inspection

Consistency Inspection is a general method for evaluating a product for consistency across interfaces. Aspects that are checked include fonts, colors, terminology, layout, input formats, interaction methods. Also the product documentation and online help is checked. Consistency inspection can be performed by a small team (2-3). C.I. can be performed at all stages of the development cycle (mock-ups, specs, prototypes).

Procedure: In advance, the style guideline (the "gold standard") against which should be checked and the attributes that should be subjected to inspection have to be defined. Then the interfaces, or parts which should be checked (i.e. mobile web site and web site) have to be selected. The inspectors are further given a template for reporting inconsistencies before they perform the inspection. Individual results are discussed, aggregated and prioritized in the team.

Outcome:

- a prioritized list of inconsistencies

Chapter HCI:IV

III. Usability Testing

- Overview
- Usability Labs
- Testing humans
- Steps
- Methods
  - Thinking aloud
  - Co-Discovery
  - Formal Experiment
  - A/B Testing
- Limitations

Usability Testing

Overview

Usability Testing: Empirical methods of interface testing with representative users.In the test users try to complete typical tasks while evaluator watches.

Methods

- **Thinking aloud:** Test users speak what they think during experiment.
- **Co-Discovery:** Two test users perform tasks together and talk to each other.
- **A/B test:** Two groups of users perform tasks with two different interface versions (e.g., current and new). Between subjects-design.
- **Formal experiment:** Controlled experiment with users, collecting measures and performing statistical analysis.

Usability Testing

Overview

Spectrum of Testing

- Test often and at different times of the software design cycle
  - ✓ Most serious problems can be found with ≈ 6 testers
  - ○ Not feasible for complex systems
- Testing to explore design alternatives or validate existing designs
  - – Compare to prior versions or to competitors' versions (A/B testing)
  - ○ Danger of experimenter bias
- Testing paper prototypes
- Remote testing (web-based applications)
  - – Recruited in online communities, via email, Amazon Mechanical Turk
  - ✓ Large number of participant
  - ○ Difficult in preparation, logging, and controlling conditions
- Field tests and portable tests
- Can-you-break-this tests
- Eye tracking (where participants are looking at and how long)

Usability Testing

Usability Labs

- Dedicated usability labs, divided by half-silvered mirror for observation
- Ad-hoc settings in offices
- Quiet room, each participant same conditions (do not change hardware during testing)
- Video recordings are advisable, participants may be uncomfortable at first, but tend to forget the recording during their tasks

Usability Lab ©http://www.cure.at/controlrooms

Simpler Setting [Granitzer2004]

Usability Testing

Testing humans

- Treat participants with respect
- Collect signed informed consent statement

Content of informed consent statement

- The purpose of the study (why is the study being done)
- If there are audio/video recordings, who will see the recordings and what happens to the material after the study is finished.
- Statement of confidentiality and how anonymity will be preserved.
- Participation is voluntary, and participants can stop at each time.
- Whom to contact in case of questions.
- The fact that all questions arisen at the beginning of the study have been answered.

Usability Testing

Testing humans

Consider how these factors might influence the results:

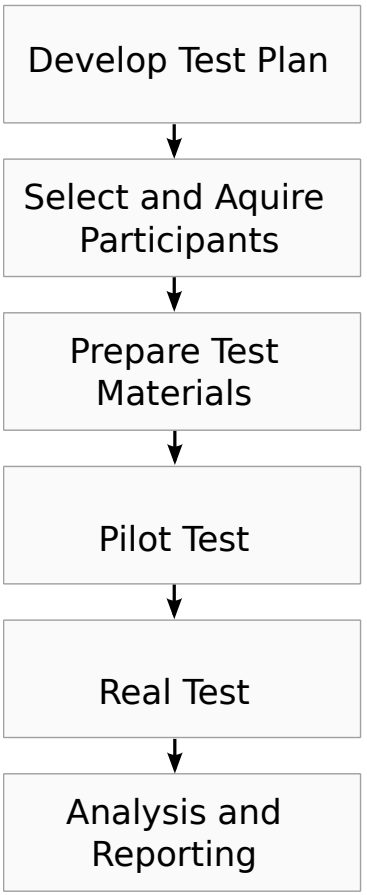
- People get tired
- People get bored
- People learn during the experiment
- People have emotions (might get upset)

Tasks should not be too long. Counterbalance interface components (first user gets condition A, then B, second user B then A and so on). Inform them that the interface is tested, not the user.

Usability Testing

Steps

- Usability experts participate in early task analysis or design reviews and set up tasks for usability tests
- Participants are defined and recruited
- Finished detailed test plan (list of tasks, questionnaire) 2-6 weeks before usability test
- Pilot testing with 2-3 users to find potential test problems (as if it were the real test, but results not integrated in final test results)
- Test is performed and results are evaluated
- A usability report is written



Usability Testing

Methods – Thinking aloud

Thinking aloud: Method where test users talk while performing the test.

What should users talk about?

- What they are trying to achieve
- What they read
- Questions that arise
- Things they find confusing or annoying
- Decisions they make

What information you get

- Finds usability problems and why they occur
- Generates nice quotes
- Users constantly read “perimeter” instead of “parameter” – not detectable when only watching people [Lewis1993]

Usability Testing

Methods – Thinking aloud

Preparation

- Demonstrate for unrelated task (e.g. booking a train ticket) or show users videos of thinking aloud tests.
- Let them practice on unrelated task and interface.
- Explain that questions should be asked during the task, but will not be answered.

During the test

- If user stops talking, encourage them to continue with neutral prompts:
  - “Can you tell me what you are doing?”
  - “Can you tell me what you are thinking now?”
- Avoid “Why” questions (you get plausible but unreliable answers)
  - Market survey showed [Nisbett1997] most people preferred rightmost pair of (3 identical) pairs of underwear. Users made up plausible reasons, real reason: natural bias towards last of similar alternatives.

Usability Testing

Methods – Thinking Aloud

Summary

- ✓ Find many usability problems and why they occur
- ✓ Require small number of users only (3-6)
- ✓ Can be performed early in design process
- ✓ Colorful quotes
  - ○ Can change user behavior (think first..)
  - ○ Slows users down (approx. 17%)
  - ○ Does not provide performance data

Example: Paper prototype usability test

http://www.youtube.com/watch?v=ppnRQD06ggY

Usability Testing

Methods – Co-Discovery

Co-Discovery: Method where two test users simultaneously explore an interface.

- Natural communication between the two users
- Similar to thinking aloud, without having to (unnaturally) talk to oneself
- Needs twice as many users
- Unnatural setting (unlikely to happen in practice)

Usability Testing

Methods – Formal Experiment

Formal Experiment: Controlled experiment with test users involving statistical analysis.

- For getting objective measurements of performance for a fully implemented design (summative evaluation).
- Use cases
  - Testing absolute performance of one interface

  $\Rightarrow$ "Users can perform Task X in 4 minutes $\pm$ 20 seconds."
  - Testing alternative designs (A/B testing).

  $\Rightarrow$ "Users perform Task X significantly faster with system A ($p = 0.05$)".

Usability Testing

Methods – Formal Experiment

Performance Measures: Objective, quantitative data

Examples:

- Task success rate (number of completed tasks within given time)
- Task completion time (time to complete task)
- Task error rates
- Feature coverage (ratio of features used to total number of features)
- Feature retention (number of features remembered after test)
- Deviation from optimal execution path (number of clicks, actions)
- Number of times help system used, time spent using help
- Ratio of positive and negative comments

Usability Testing

Methods – Formal Experiment

Testing multiple conditions

Within-subjects design: also called repeated measures design

- ❑ All participants get the same conditions
- ✓ need fewer participants
- ○ carryover effects (remembering, performing first condition influences performance on second condition, learning)

Between-subjects design: also called between-groups design

- ❑ Participants are split into groups, each group is tested on one condition solely
- ○ Requires more participant for significant results
- ○ Danger of bias (assignment bias, observer-expectancy, participant-expectancy)

⇒ Random assignments of participants to groups

⇒ Double blind studies (not feasible for UI experiments)

Usability Testing

Methods – Formal Experiment

Hypotheses

Testable statements involving dependent and independent variables

- Independent variables: conditions that change in experiment (i.e. input method, UI design)
- Dependent variables: measures that depend on the independent variables
- Hypotheses can not be “accepted” with statistical tests!
- If your assumption is that interface A is better than interface B you state your Null-Hypothesis

  H_0 *Users perform tasks X in the same amount of time using interface A and interface B.*

- Possible outcomes:
  - H_0 must be rejected. Then use the difference of mean task completion values to show that indeed A is better.
  - H_0 can not be rejected (which does not mean it is true).

Usability Testing

Methods – Formal Experiment

Summary

- ✓ Obtain objective, quantitative data (performance measures)
- ✓ Allows to compare alternative designs
- ○ Needs significant number of test users (> 20) for statistically significant results
- ○ Requires sound statistical knowledge of evaluators
- ○ Do not convey “why” something happened

Usability Testing

Methods – Formal Experiment

Statistical Analysis

How many test users?

- Rule of thumb: 20 test users (e.g. T-Test not applicable for samples $n < 20$)
- To detect smaller differences, 50-100 users might be necessary.
- The more people the more representative your study is.

Test validity – Measures the relevance of the data to the real world. Take care to

- Test with the right kind of users
- Test the right tasks
- Set up appropriate test environment (similar to practical work environment)

Usability Testing

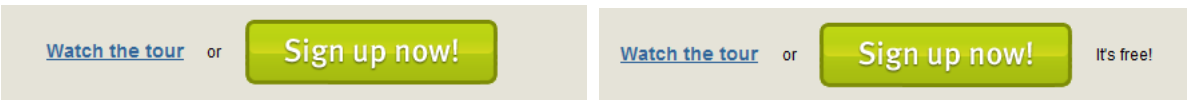
Methods – A/B Testing

A/B Testing: Formal experiment in web setting with real, live users comparing different conditions.

- One group of visitors is assigned user interface design A and one design B (only slight differences)
- Cookies make sure that each users gets the constantly the same condition.
- Dependent variables are measured (e.g., how often clicked on ads) and statistically compared across the variants.
- Constantly performed by Google, Amazon, Zynga

Example

- Soocial improved their click-through rate from 14.5% to 18.6% by adding "It's free" to the sign-in button.[1]



[1] visualwebsiteoptimizer.com/split-testing-blog/ab-test-case-study-how-two-magical-words-increased-conversion-rate-by-28/

Usability Testing

Methods – Summary

- Effectiveness: Capability of producing desired result (task completion rate)
- Efficiency: Extend of well-usage of resources (task completion time)
- Satisfaction: Subjective measure of “happiness

	Thinking-Aloud	Co-Discovery	Formal Exp.	A/B testing
Effectiveness	✓	✓	✓	✓
Efficiency	-	-	✓	-
Satisfaction	✓	✓	✓	-
# Users	3-6	6-12	20/100	thousands

Usability Testing

Limitations

Emphasizes first time usage

- Usability tests are usually 1-3 hours, difficult to estimate the performance after weeks or months of usage

Provides limited coverage of the interface components

- During the short usability test, participants work only with a limited set of features and/or interface components

Non-realistic test environment

- Longer term testing might be necessary to understand adoption and learning processes
- Realistic tests of interfaces in mission-critical systems (military combat) or high-stress situations (surgery) can not be done in the usability lab

Usability Testing of Fruit:
https://www.youtube.com/watch?v=3Qg80qTfzgU

Chapter HCI:IV

IV. Survey Techniques

- Overview
- Preparing a Survey
- Sample Questionnaires

Survey Techniques

Overview

Questionnaires and Surveys: are familiar and inexpensive addition for usability tests and expert reviews.

- managers and users can immediately grasp the notion of surveys
- can be cheaply performed with a large number (hundreds to thousends) of users
- less specific and controlled than expert reviews and usability tests
- far more coverage across user population than the latter two leading to more impressive statements (nearly 70% of the 500 testers prefered interfac A to interface B)

Survey Techniques

Preparing a Survey

Procedure:

1. Prepare survey form, review with experts (colleagues)
2. Test on small sample of users
3. Prepare methods of statistical analysis (statistical tests) and presentation (histograms, scatterplots)
4. Prepare distribution process to collect surveys from representative users (age, gender, experience, ..)
5. Make sure you collect only one response per user and anonymize the data (unique user ID for users sent by e-mail, mapping e-mail address user ID is not stored)
6. Distribute the survey
7. Evaluate the results

Survey Techniques

Preparing a Survey

Content:

- Subjective impression of user interface (design, interaction, understandability)
- User Characteristics
  - Background Demographics (age, gender, origin, income, native language, education)
  - Experience with Computers (specific software, duration of use, depth of knowledge)
  - Job Responsibilities (decision-makers, managing)
  - Familiarity with features (do you know the “printing” function)
  - Emotional state after using an interface (frustrated, confused, happy,..)

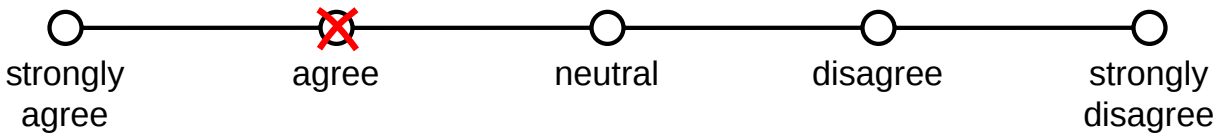
When using web-based survey (which are very cheap compared to paper) one must be aware of the fact that only web-affine persons will take part (bias in population).

Survey Techniques

Designing Survey Questions

Likert Scale [Likert1932]

1. Wikipedia has a user friendly interface.



©Wikipedia

Bipolar Semantically Anchored Scales [Coleman1985]

hostile	1	2	3	4	5	6	7	friendly
vague	1	2	3	4	5	6	7	specific

General Advice

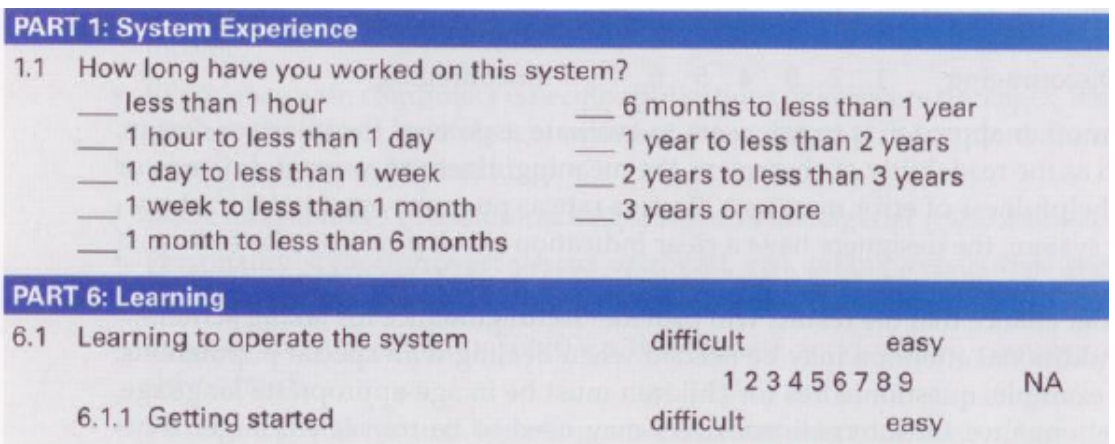
- Avoid leading questions, because they bias the answers. (e.g. “Would you like an interface with less clutter?”)
- Language must be appropriate (children vs. adults, technical computer users vs. novices)
- Be specific (about the component to rate, the dimension that you ask for) to have useful results for guiding the subsequent redesign process.

Survey Techniques

Sample Questionnaires

Questionnaire for User Satisfaction (QUIS) `http://lap.umd.edu/quis/`

- a demographic questionnaire
- measure of overall system satisfaction
- hierarchically organized measures of nine specific interface factors (e.g., learning factors, screen factors)
- needs to be licensed, on-line preview (`http://lap.umd.edu/quis_net/`)



PART 1: System Experience

1.1 How long have you worked on this system?

___ less than 1 hour	___ 6 months to less than 1 year
___ 1 hour to less than 1 day	___ 1 year to less than 2 years
___ 1 day to less than 1 week	___ 2 years to less than 3 years
___ 1 week to less than 1 month	___ 3 years or more
___ 1 month to less than 6 months	

PART 6: Learning

6.1	Learning to operate the system	difficult 1 2 3 4 5 6 7 8 9 easy	NA
	6.1.1 Getting started	difficult easy	

[Shneiderman & Plaissant 2010]

Survey Techniques

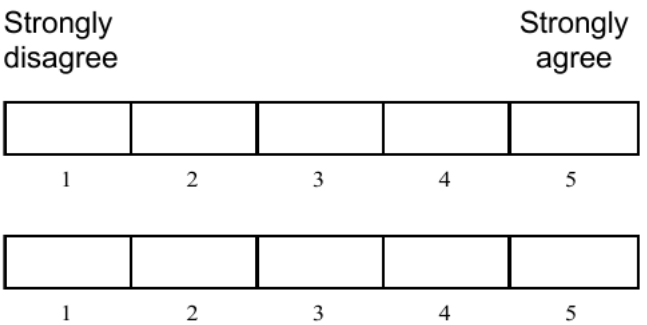
Sample Questionnaires

System Usability Scale (SUS) [Brooke1996]

- 10 statements with which users rate their agreement
- Scale is always the same (strongly disagree – strongly agree)
- “Quick and dirty” method for low cost assessment of usability
- Half of the questions positively worded, half negatively

1. I think that I would like to use this system frequently

2. I found the system unnecessarily complex



[Brooke1996]

Survey Techniques

Sample Questionnaires

Post-Study Usability Scale (PSSUQ) [Lewis1995]

- Developed by IBM
- 48 statements with which users rate their agreement
- Scale is always the same (strongly disagree – strongly agree)
- Focus on overall satisfaction, system usefulness, information quality and interface quality

1. Overall, I am satisfied with how easy it is to use this system.

STRONGLY AGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**STRONGLY DISAGREE**

COMMENTS:

2. It was simple to use this system.

STRONGLY AGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**STRONGLY DISAGREE**

COMMENTS:

[Lewis1995]

Survey Techniques

Sample Questionnaires

Software Usability Measurement Inventory (SUMI)

http://sumi.ucc.ie/en/

- available in different languages (including German)
- 50 statements to measure emotional response, efficiency, control, learnability and helpfulness of the interface
- scale is always the same (agree, undecided agree)
- needs to be licensed, academic license available

Statements 1 - 10 of 50.	**Agree**	**Undecided**	**Disagree**
This software responds too slowly to inputs.	○	○	○
I would recommend this software to my colleagues.	○	○	○

[http://sumi.ucc.ie/en/]

Survey Techniques

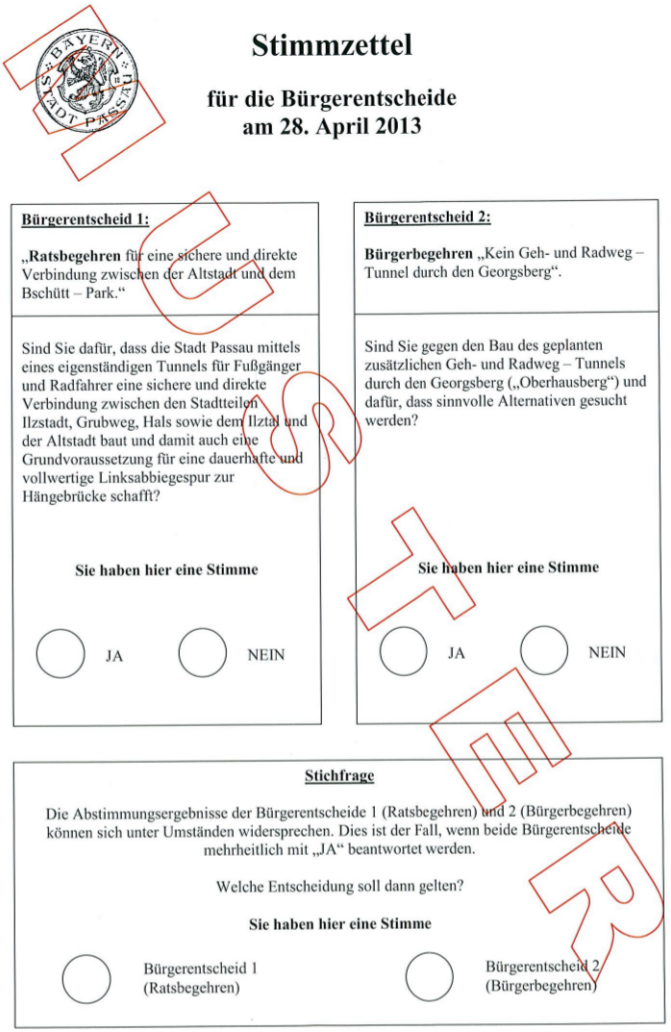
Sample Questionnaires

Others

- Computer Systems Usability Questionnaire (CSUQ)
- Website Analysis and Measurement Inventory (WAMMI) questionnaire
- Mobile Phone Usability Questionnaire (MPUQ)

Survey Techniques

Counter-Example



Stimmzettel

für die Bürgerentscheide am 28. April 2013

Bürgerentscheid 1:

„**Ratsbegehren** für eine sichere und direkte Verbindung zwischen der Altstadt und dem Bschütt – Park."

Sind Sie dafür, dass die Stadt Passau mittels eines eigenständigen Tunnels für Fußgänger und Radfahrer eine sichere und direkte Verbindung zwischen den Stadtteilen Ilzstadt, Grubweg, Hals sowie dem Ilztal und der Altstadt baut und damit auch eine Grundvoraussetzung für eine dauerhafte und vollwertige Linksabbiegespur zur Hängebrücke schafft?

Sie haben hier eine Stimme

JA NEIN

Bürgerentscheid 2:

Bürgerbegehren „Kein Geh- und Radweg – Tunnel durch den Georgsberg".

Sind Sie gegen den Bau des geplanten zusätzlichen Geh- und Radweg – Tunnels durch den Georgsberg („Oberhausberg") und dafür, dass sinnvolle Alternativen gesucht werden?

Sie haben hier eine Stimme

JA NEIN

Stichfrage

Die Abstimmungsergebnisse der Bürgerentscheide 1 (Ratsbegehren) und 2 (Bürgerbegehren) können sich unter Umständen widersprechen. Dies ist der Fall, wenn beide Bürgerentscheide mehrheitlich mit „JA" beantwortet werden.

Welche Entscheidung soll dann gelten?

Sie haben hier eine Stimme

Bürgerentscheid 1 (Ratsbegehren)

Bürgerentscheid 2 (Bürgerbegehren)

Chapter HCI:IV

V. Acceptance Tests

- Overview
- Example Specifications
- Benefits and Limitations

Acceptance Tests

Overview

Acceptance Tests: Testing measurable criteria of the user interface as specified in the requirements document.

Measurable criteria of user interfaces:

- Time to learn specific functions
- Speed of task performance
- Task completion
- Rate of errors made by users
- User retention of commands over time
- Subjective user satisfaction
- System response time

Acceptance Tests

Example Requirements

Specified in requirements document

- Task Completion

  The participants will be 35 adults (25-45 years old), native speakers with no disabilities, hired from an employment agency. They will have moderate web-use experience: 1-5 hours/week for at least a year. They will be given a 5-minute demonstration of the basic features. At least 30 of 35 adults should be able to complete the benchmark task within 30 minutes.

- User Retention

  Ten participants will be recalled after one week and asked to carry out a new set of benchmark tests. In 20 minutes, at least eight of the participants should be able to complete the task correctly.

Acceptance Tests

Benefits and Limitations

- Help to avoid arguments about user friendliness
- Contract fulfillment can be objectively demonstrated
- Central goal is to verify compliance with requirements documents
- Can not to detect flaws as (usability tests or expert reviews)
- Performed in pre-release phase, when changes are relatively cheap

Chapter HCI:IV

VI. Evaluation During Active Use

- Interviews and Focus-Group Discussions
- Continuous Data Logging
- Help-Desks and Internet Communities
- Automatic Evaluation

Evaluation During Active Use

Interviews and Focus-Group Discussions

Help to identify very focused issues (**individual interviews**) and ascertain the universality of comments (**focus-group discussions**).

Example: 66 out of 4300 users of internal messaging were interviewed in a large company, each interview 45'.

- ✓ Users happy with capacity to pick-up messages anywhere
- ✓ Users happy with legibility of printed messages
- ✓ Users happy with convenience of after-hour access
- – 23.6% of users concerned about reliability
- – 20.2% found using the system confusing
- – 23.6% found that convenience and accessibility could be improved
- ✓ Only 16% no concerns

This result lead to 42 enhancements of the interface reflecting the users' needs.

Evaluation During Active Use

Continuous Data Logging

Continuous Data Logging: refers to software-based data logging including interface usage, speed of user performance, rate of errors and requests for assistance. Is extensible and commercially used in web context (web analytics).

Potential results:

- Find most frequent error and react
- Not occurring errors may indicate unused functionality
- Direction for interface changes (make most used functionality highly efficient to use)
- Guidance for acquiring new hardware (data base server, webserver)
- Devising training programs
- Finding and fixing (usability) bugs

The fact that nearly everything can be logged does not necessarily mean that it should be. **Take care of privacy issues (and on country-specify privacy laws).**

Evaluation During Active Use

Help Desks, Internet Communities

Help Desk systems include telephone help lines, special email addresses (support@...) and on-line feedback tools (e.g., bugzilla).

- Users feel reassured if there is a human who can help them if need be
- Telephone help lines may also have remote desktop access and can user walk through the solution (efficient communication, learning effect for users, but not always feasible)

Internet Communities include discussion groups, wikis and newsgroups. Help is given from users to users. Conversation can be followed and crucial issues can be detected.

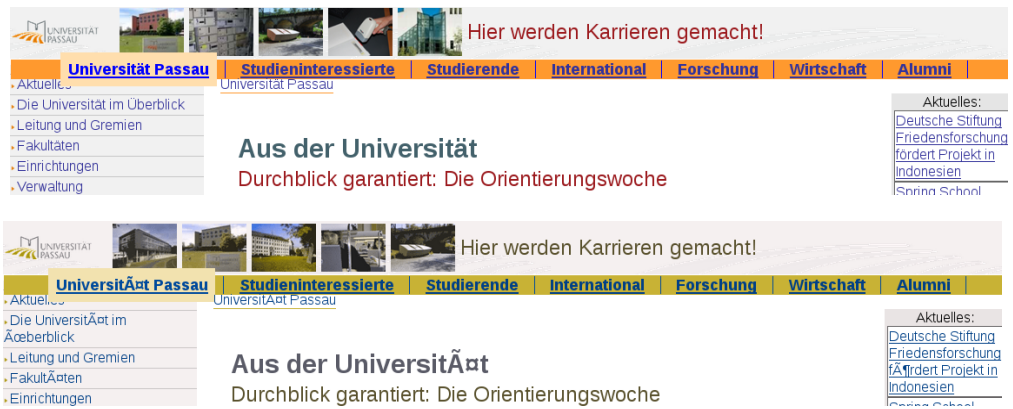
Evaluation During Active Use

Automatic Evaluation

Automatic Evaluation is performed by software tools, automatically checking interfaces against a set of guidelines.

Examples:

- WWW markup validation[2] for HTML, XML ("you forgot the alt-tag")
- Website validation for color perception[3]
- Website layout metrics (informational, navigational and graphical aspects)



Uni Passau Homepage as perceived by users with normal vision (top) and red-green color blindness (bottom)

[2] http://validator.w3.org
[3] http://colorfilter.wickline.org/

Chapter HCI:IV

VII. Example User Study

Example User Study

Overview

Background:

- Example user study taken from [Seifert2013]
- Task: Categorization of news articles into “politics”, “science”, “economy”, “sports”, “culture”
- Needs to be done manually (at least to generate training data for supervised classifiers), time-consuming work

Question:

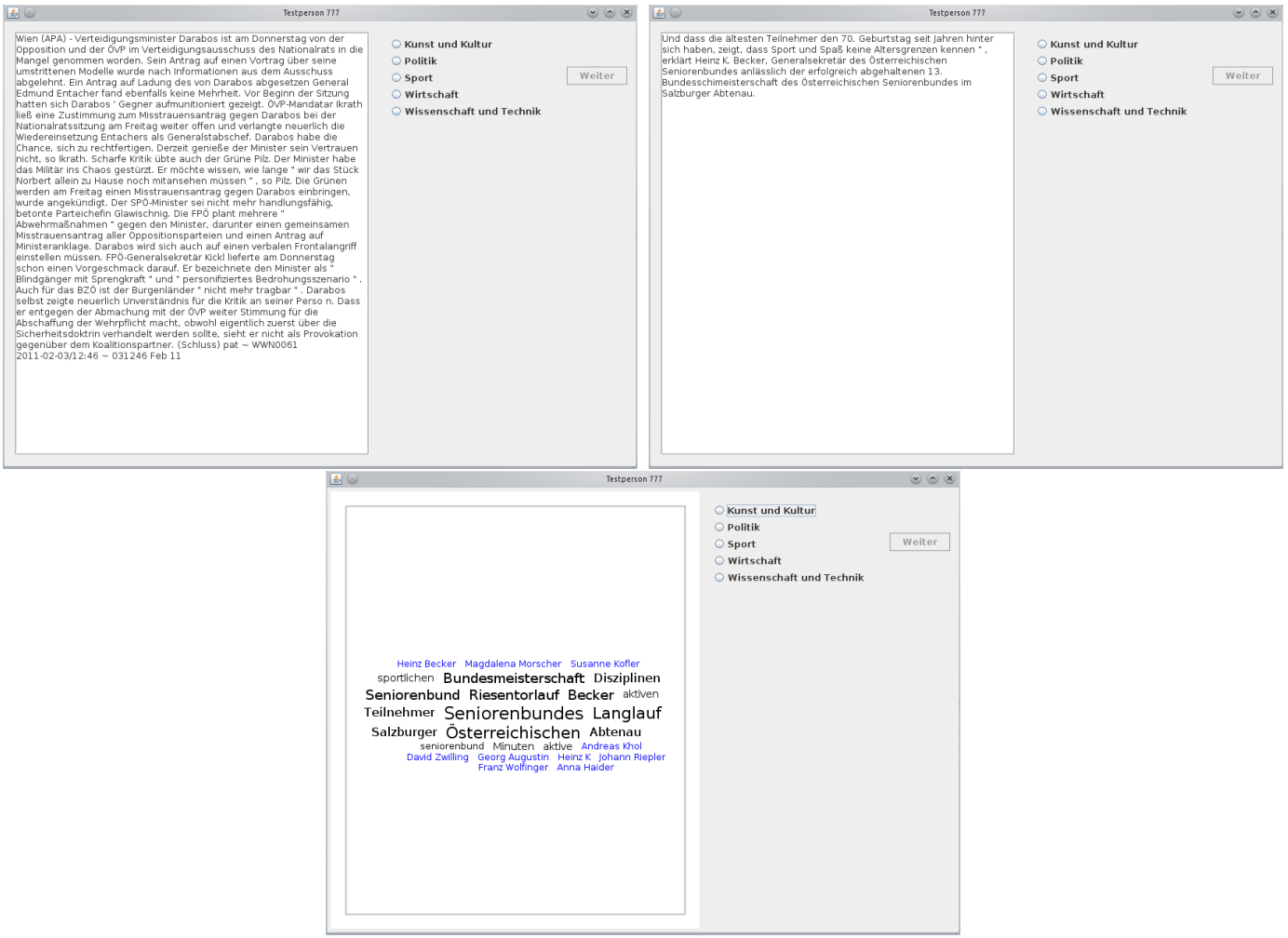
Can we design a visualizatin of the news articles that allows user to perform the task faster (while retaining accuracy)?

Idea:

Use different versions of text summaries (full-text, automatically extracted key sentences, automatically extracted keywords).

Example User Study

GUI



Example User Study

Settings

Hypotheses

H1 The time required for labeling key phrases or key sentences is significantly less than for labeling full-text documents

H2 There is no difference in the number of correct labels between key phrases, key sentences and full-text.

Participants

- 37 German speaking, 19 males, 18 females
- age 25 to 58 years
- 23 technical professionals (including students)

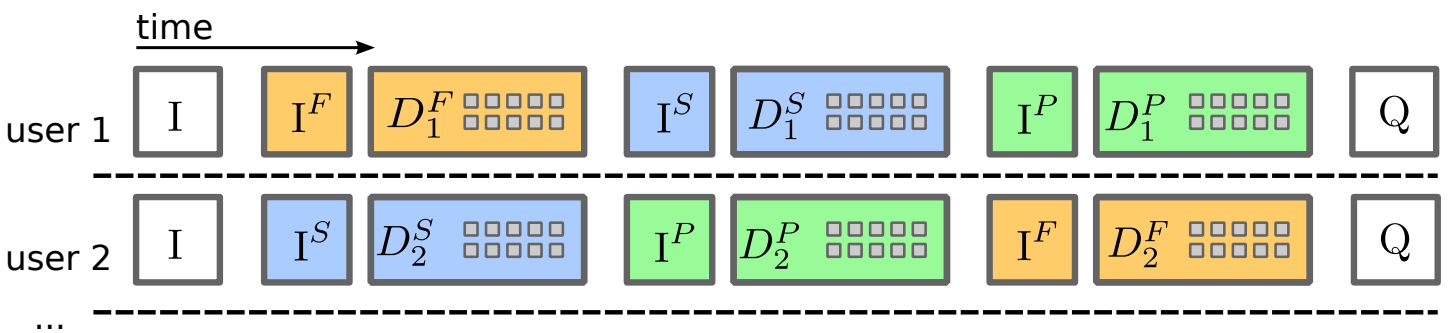
Test Material

- German news corpus, articles of 2008
- 27570 news articles, $2 \leq words \leq 2720$
- chose articles with length in 3rd quartile $\rightarrow$ 6328 news articles
- fully labeled, 5 classes: economy, politics, sports, culture, science

Example User Study

Evaluation Procedure

- Begin with general introduction (I), get informed consent statement signed
- Within-subjects design, sequence of conditions (D^F, D^S, D^P) for users shuffled
- Introduce task at the beginning of each condition (I^F, I^S, I^P)
- End with questionnaire (Q)



Example User Evaluation Procedure

Example User Study

Results

- 370 labeled documents per condition, correct
  - full-text: 290
  - key sentences: 281
  - key words: 305

Table: Overview of labeling time and number of correct labels (out of 10) for each condition. Values averaged over all users, showing mean and standard deviation.

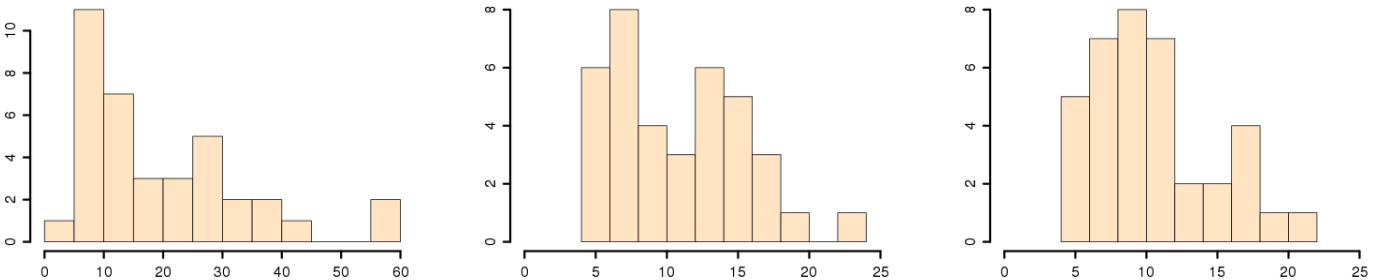
	full-text	**key sentences**	**key phrases**
correct labels	7.84 ± 1.24	7.59 ± 1.38	$\mathbf{8.24 \pm 1.23}$
completion time [s]	19.9 ± 13.8	10.7 ± 4.4	$\mathbf{10.4 \pm 4.1}$

Differences significant?

Example User Study

Results

H1 - Labeling time



- H_0 for test: no difference in mean values.
- No normal distribution, compare means with Wilcoxon rank sum test for unpaired samples, $\alpha = 0.05$
- Conclusion for H1: **Full-text slower than key sentences and slower than key phrases; no difference between key phrases and key sentences**
- Conclusion for H2[4]: **Key sentences less accurate than full-text and key phrases; no difference between key phrases and full-text**

[4] not shown in detail here

Example User Study

Performance vs. subjective measures

- User do not necessarily perform the way they feel they perform
- There is a difference between user performance and subjective performance (which is related to user satisfaction), neither of which can be neglected
- Measured Performance (time to the next click, accuracy of results against ground truth, the more labels the better, the less time the better)

	full-text	**key sentences**	**key phrases**
labels	7.84 ± 1.24	7.59 ± 1.38	$\mathbf{8.24 \pm 1.23}$
time [s]	19.9 ± 13.8	10.7 ± 4.4	$\mathbf{10.4 \pm 4.1}$

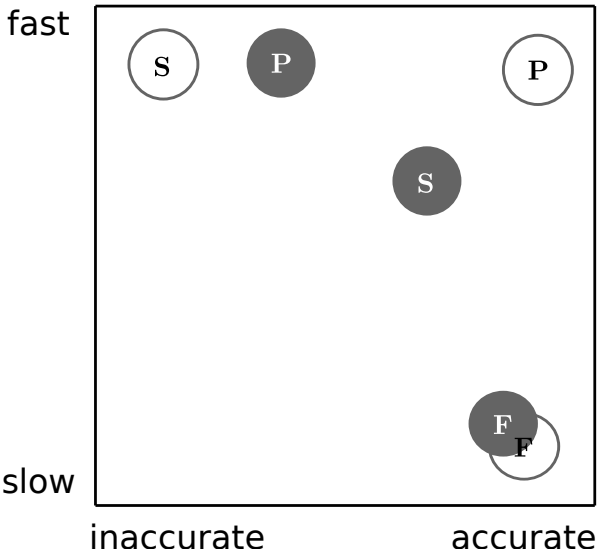
- Perceived performance (questionnaire results)

question	**full-text**	**key sentences**	**key phrases**
speed	$3.6 \pm 1.0^*$	3.9 ± 1.0	$\mathbf{4.0 \pm 1.0}$
difficulty	$\mathbf{4.1 \pm 1.0}$	3.8 ± 1.2	3.4 ± 1.2
enough hints	$\mathbf{4.7 \pm 0.8}$	3.8 ± 1.1	3.2 ± 1.1

Example User Study

Performance vs. subjective measures

- Filled circles – perceived performance, empty circles – measured performance
  1. in some conditions users feel fast, but are slower (S)
  2. in some conditions users feel inaccurate, but are very accurate (P)
- In this case the cause is likely that one tested condition (P) was new to users, thus they felt less comfortable (but nevertheless performed measurable better)



Perceived vs measured performance for 3 conditions S, P, and F

Make sure you test what you want to find out

Chapter HCI:IV

VIII. Summary

Summary

Evaluation Methods

- **Expert Reviews:** the evaluator is part of the product team (colleagues, consultants), not the actual user
- **Usability Testing:** formal testing and observation of users in controlled environment
- **Surveys:** questionnaires for users
- **Acceptance Tests:** testing against requirements, outside lab environment
- **Evaluation During Active Use:** interviews, data logging, after product shipment

Most important

- Always perform a pre-study
- Always test